A Generic Machine Learning Framework for Fully-Unsupervised Anomaly Detection with Contaminated Data

Markus Ulmer[1], Jannik Zgraggen[2], and Lilach Goren Huber[3]

[1,2,3] *Zurich University of Applied Sciences, Winterthur 8401, Switzerland*
markus.ulmer@zhaw.ch
jannik.zgraggen@zhaw.ch
lilach.gorenhuber@zhaw.ch

ABSTRACT

Anomaly detection (AD) tasks have been solved using machine learning algorithms in various domains and applications. The great majority of these algorithms use normal data to train a residual-based model and assign anomaly scores to unseen samples based on their dissimilarity with the learned normal regime. The underlying assumption of these approaches is that anomaly-free data is available for training. This is, however, often not the case in real-world operational settings, where the training data may be contaminated with an unknown fraction of abnormal samples. Training with contaminated data, in turn, inevitably leads to a deteriorated AD performance of the residual-based algorithms.

In this paper we introduce a framework for a fully unsupervised refinement of contaminated training data for AD tasks. The framework is generic and can be applied to any residual-based machine learning model. We demonstrate the application of the framework to two public datasets of multivariate time series machine data from different application fields. We show its clear superiority over the naive approach of training with contaminated data without refinement. Moreover, we compare it to the ideal, unrealistic reference in which anomaly-free data would be available for training. The method is based on evaluating the contribution of individual samples to the generalization ability of a given model, and contrasting the contribution of anomalies with the one of normal samples. As a result, the proposed approach is comparable to, and often outperforms training with normal samples only.

1. INTRODUCTION

Anomaly detection (AD) tasks are common in very diverse fields, including medical image processing, autonomous driving, fraud detection, and fault detection in industrial machines. An inherent property of AD tasks is that very few or no labeled examples of anomalous behavior are provided in advance.

Therefore, the most common machine learning approaches to solve these tasks are based on using exclusively normal data to train a selected prediction algorithm, which is subsequently used to infer on unseen data. The underlying assumption here is that the algorithm's prediction errors (residuals) will be higher whenever the input sample does not belong to the learned distribution. This family of models can be referred to as residual-based models, irrespective of whether they use regression or reconstruction residuals to detect anomalies, with the most common models being reconstruction models such as principal component analysis (PCA) or various types of autoencoder (AE) neural networks. It is worth noting that these models are often termed "unsupervised". However, in the context of AD, they should be referred to as "semi supervised" since they assume the availability of labeled normal data for training.

In fact, such models tend to perform rather poorly whenever contamination in the form of anomalous samples is introduced into the training data. In real-world applications, however, the assumption of having anomaly-free training data does not always hold, as data contamination cannot be avoided. In this case, truly unsupervised methods, based on clustering or one-class classification (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 1999), are required. Recently there has been a growing effort to develop deep unsupervised algorithms for AD, whose performance is not severely damaged by data contamination (Munir, Siddiqui, Dengel, & Ahmed, 2018).

Despite the high practical relevance of the problem, systematic solutions for AD with contaminated training data are rather rare. Recently, several papers have suggested useful approaches based on data refinement (Yoon et al., 2021), on latent outlier exposure (Qiu, Li, Kloft, Rudolph, & Mandt, 2022), and on physics-informed deep learning (Zgraggen, Guo, Notariste-

https://doi.org/10.36001/IJPHM.2024.v15i1.3589

fano, & Goren Huber, 2023).

In this paper, we propose a novel framework that addresses the challenge of AD with potentially contaminated training data. Our approach here is generic: the framework can be used with any residual-based machine learning model (e.g. PCA, any kind of AE or regression NN). Our aim is to offer a substantial improvement of the residual-based approach by allowing any residual-based AD model (intended to be trained on anomaly-free data) to be applicable in a fully unsupervised setting without any labels and not assuming anomaly-free training data.

The framework is based on a single-step (not iterative) refinement of the contaminated data which proposes candidate anomalies to be removed from the training data. The refinement algorithm itself does not depend on the contamination ratio. The only assumption behind it is that the majority class is the normal one (which is a defining property of any AD task).

We demonstrate the performance of the proposed framework on two industrial time series datasets. The first dataset is of high-frequency acoustic sensor data converted to Mel Spectrograms. With this dataset, we address the use-case of abrupt machine faults of various types. The second dataset is a multivariate time series from Turbofan aircraft engines aimed at monitoring the gradual degradation of the engines over their lifetime.

For each dataset, we evaluate the data refinement quality by comparing to the naive approach of training a residual-based model blindly with the entire contaminated data. For reference, we also compare the performance to the ideal (but rather unrealistic) case of training with anomaly-free data. We show that the refinement step is essential in order to achieve high performance AD in the presence of significant contamination.

The contribution of the paper is in addressing the highly relevant challenge of AD under realistic conditions in which anomaly-free training data is not available. The approach we take is simple for implementation with any existing residual-based AD model, and can be applied on raw data or on learnable representations. Yet, it is shown to perform as well as the ideal reference of training the same model with anomaly-free data and sometimes outperforms it in its refinement efficacy, as it utilizes the properties of the anomalous samples and contrasts them with the normal samples. The focus of this paper is on the effectiveness of the framework for AD on time series data. However, the framework is generic in nature and should in principle be applicable to any data type.

2. RELATED WORK

AD with machine learning. A large variety of machine learning methods for AD have been developed in recent years. The great majority of the work has focused on standard AD problems, in which the training data is assumed to be anomaly-free. Many of the classical algorithms are distance or density-based, like one-class classifiers (Schölkopf et al., 1999; Tax & Duin, 2004) or density estimation methods (Latecki, Lazarevic, & Pokrajac, 2007). Later works suggest various deep learning models, mostly based on autoencoders (AEs) that are trained on normal data and detect anomalies based on the model residuals at inference. Such models have been applied to various data types including images (Zhou & Paffenroth, 2017) and multi-variate time series (Audibert, Michiardi, Guyard, Marti, & Zuluaga, 2020; Munir et al., 2018; Zhang et al., 2019).

More recently it has been demonstrated that state-of-the-art AD performance can be achieved by one-class classifiers on pre-trained features extracted from deep learning architectures (Sohn, Li, Yoon, Jin, & Pfister, 2020). Different variants of self-supervised feature extractors have been exploited for AD on image data (Golan & El-Yaniv, 2018; Bergman & Hoshen, 2020; Hendrycks, Mazeika, & Dietterich, 2018) as well as on tabular and time series data (Shenkar & Wolf, 2022; Schneider et al., 2022; Michau, Frusque, & Fink, 2022).

As explained above, the common prerequisite to all of the above methods is anomaly-free ("normal") training data. Since this assumption is rarely valid in practical settings, the present paper deals with the challenge of a contaminated training dataset.

AD with contaminated training data. The most common approach to address the problem of anomaly-contamination in the training data is to assume that the fraction of anomalies is low enough, so that the standard AD algorithms can be "blindly" trained on the contaminated data (Zong et al., 2018; Bergman & Hoshen, 2020). However, as has been shown before, this assumption breaks down already for rather low contamination ratios (e.g 5% with robust approaches (Munir et al., 2018)).

An alternative approach to deal with contaminated data are iterative refinement methods, which use one-class classifiers (OCCs) (Beggel, Pfeiffer, & Bischl, 2020) or reconstruction errors of the AE (Zhou & Paffenroth, 2017; Berg, Ahlberg, & Felsberg, 2019) in order to remove the anomalies and improve the AD performance subsequently. (Yoon et al., 2021) suggests to boost the refinement performance yet more using an ensemble of OCCs instead of a single model. Recent papers (Qiu et al., 2022; Wang, Zhan, Wang, Song, & Nahrstedt, 2022) suggest that contrasting information from the anomalous class helps to improve the AD performance with contaminated training data.

Our approach makes conceptual use of ideas of the above papers, but suggests a new simple approach, inspired by data centric concepts. The proposed framework refines the con-

taminated training data by splitting it into partially overlapping subsets and training an ensemble of residual based models. The refinement score assigned to each sample contains information about its contribution to the generalization ability of the trained model. This is done by contrasting the AD performance of ensemble members trained with and without this sample in the training set. Our approach differs from the ones mentioned above by its simplicity and by its generic nature. The idea can be applied with any residual-based model, be it a reconstruction model (PCA or any variant of AE) or a regression model (from linear regression to deep neural networks). The suggested framework has no assumption regarding the contamination rate, and relies solely on the basic anomaly-detection assumption of a normal majority class. Moreover, the method is demonstrated here for time series data, but should be conceptually valid for any data type. In this sense the framework is very compatible with the recent trend of data centric AI approaches.

3. METHOD

We assume a training dataset $\mathcal{D}_0 = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ containing N samples of normal data with abnormal contamination where $\mathbf{x}_i$ are the input and $\mathbf{y}_i$ the output variables. In contrast to the common notation in AD problems, here the target variables $\mathbf{y}_i$ are not the outputs of a binary classifier that can only obtain the values 0 and 1. The reason is, that our method suggests a generic framework that applies to any residual-based prediction model $\mathbf{y}_i = f(\mathbf{x}_i)$ which is traditionally trained with normal data. Note that in case of a reconstruction model (like PCA or AE), the target variables $\mathbf{y}_i$ are equal to $\mathbf{x}_i$ and the predictions are thus $\hat{\mathbf{x}}_i$. On the other hand, in a regression setup, the outputs $\mathbf{y}_i$ are typically different from the inputs $\mathbf{x}_i$. In any case, no normal/abnormal labels are assumed during the entire training and inference process, and in this sense the data is unlabeled.

3.1. Proposed Framework

The goal of the proposed method is to refine the unlabeled contaminated training data $\mathcal{D}_0$, such that at a second step only normal samples can be used to ideally train a residual-based model of choice.

The suggested USDR framework is model agnostic. Any residual-based model can be selected, and is trained with multiple subsets of the original training data $\mathcal{D}_0$. When inferring with the resulting ensemble of trained models on a sample from the original training data, we use all of the ensemble residuals to construct a "refinement score" for this sample. This score is then utilized to refine the training data, and isolate the anomalous samples. The USDR framework includes three steps:

Step I: Training a residual-based model on subsets of the training data. The unlabeled training data $\mathcal{D}_0$ is split into M equally sized *partially overlapping* subsets $\{\mathcal{D}_j\}_{j=1}^M$, such that each sample $\mathbf{x}_i$ appears in M_{train} of the M sets. The values of M and M_{train} are determined such that each subset is large enough to allow for training in the known (clean) case, and that both M_{train} and $M - M_{\text{train}}$ are large enough to allow for statistics, as explained below. Moreover, when defining the subsets, we ensure that each sample is represented in an equal number of training subsets.

In the present paper we focus on detecting contextual anomalies in time series data. In this case, a simple way to construct the training subsets $\{\mathcal{D}_j\}_{j=1}^M$ is by sliding a window of w samples with a stride of d on the entire time ordered training data. In order to make sure that every sample is equally represented in the training subsets, we use periodic boundary conditions on the training data $\mathcal{D}_0$ when constructing the overlapping subsets. The parameter d is chosen to guarantee large enough subset numbers M_{train} and $M - M_{\text{train}}$.

After obtaining the subsets $\{\mathcal{D}_j\}$, we train a selected model f on the inputs $\mathbf{x}_i \in \mathcal{D}_j$ for a given j, $1 \leq j \leq M$ to predict the targets $\mathbf{y}_i$. The model f is expected to act as a residual-based AD model when trained with normal data. However, here we train it with unlabeled data, which may contain abnormal samples. We denote with f_j the model trained with subset $\mathcal{D}_j$. After training on all subsets, we end up having an ensemble of trained models $\{f_j\}_{j=1}^M$, each of which was trained with a subset of the original unlabeled training data. We note again that a given sample $\mathbf{x}_i$ is included only in M_{train} of the M subsets, and not in all of them (see Figure 1).

Step II: Using the trained ensemble for inference on the entire data. In this step we use the trained ensemble $\{f_j\}_{j=1}^M$ to infer on the entire unlabeled dataset $\mathcal{D}_0$. We denote the prediction of model f_j using the input $\mathbf{x}_i$ by $\hat{\mathbf{y}}_{ij}$:

$$\hat{\mathbf{y}}_{ij} = f_j(\mathbf{x}_i). \tag{1}$$

Step III: Assigning a refinement score to each sample. In this step we use the predictions $\hat{\mathbf{y}}_{ij}$ to assign a refinement score S_i^{USDR} for each sample $\mathbf{x}_i \in \mathcal{D}_0$. To this end, we first separately rescale the residuals of each individual member j of the ensemble. The rescaled residual of sample i with model j is defined as:

$$r_{ij} \equiv \frac{|\mathbf{y}_i - \hat{\mathbf{y}}_{ij}| - \mu_j}{\sigma_j} \tag{2}$$

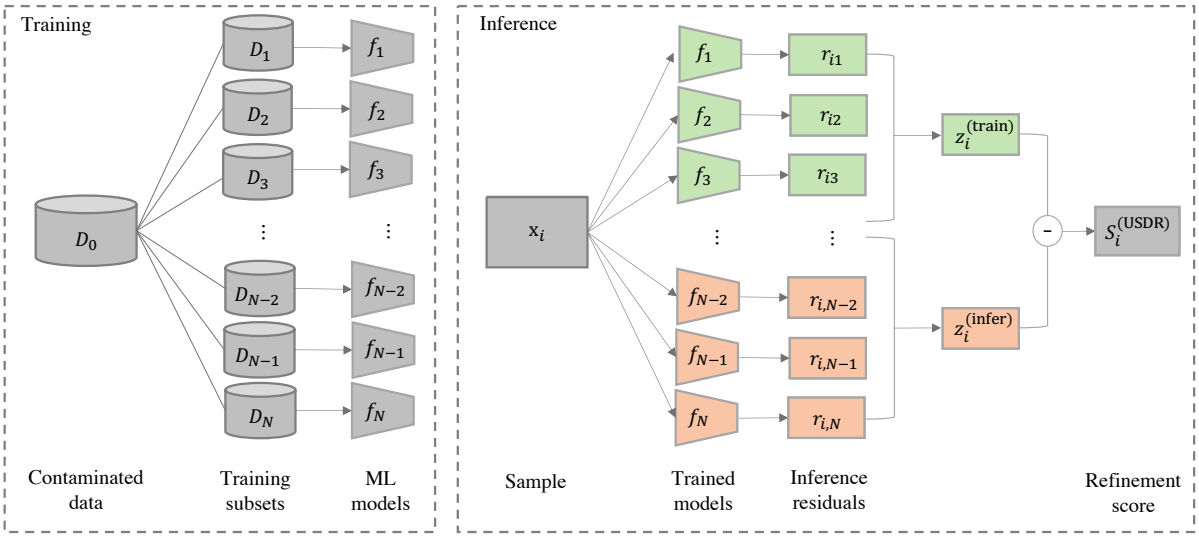

Figure 1. The proposed Unsupervised Data Refinement framework.

where

$$\mu_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in \mathcal{D}_j} |\mathbf{y}_i - \hat{\mathbf{y}}_{ij}| \tag{3}$$

$$\sigma_j^2 = \mathrm{VAR}\left(|\mathbf{y}_i - \hat{\mathbf{y}}_{ij}|\right) \tag{4}$$

are the mean and variance of the residuals within the training subset $\mathcal{D}_j$.

Next, each sample in the original training data obtains a refinement score based on its individual contribution to the training generalization ability. To this end, we define two types of residual means for each sample of the training dataset $\mathcal{D}_0$. The mean residual of sample $\mathbf{x}_i$ over all the models f_j trained with a subset $\mathcal{D}_j$ which *includes* the sample $\mathbf{x}_i$ is defined as:

$$z_i^{(\text{train})} \equiv \frac{1}{M_{\text{train}}} \sum_{j=1}^{M_{\text{train}}} r_{ij}, \mathbf{x}_i \in \mathcal{D}_j \tag{5}$$

The mean residual of sample $\mathbf{x}_i$ over all the models f_j trained with a subset $\mathcal{D}_j$ which *does not include* the data point $\mathbf{x}_i$ (i.e $\mathbf{x}_i$ is outside the training subset) is defined as:

$$z_i^{(\text{infer})} \equiv \frac{1}{M - M_{\text{train}}} \sum_{j=1}^{M-M_{\text{train}}} r_{ij}, \mathbf{x}_i \notin \mathcal{D}_j \tag{6}$$

The anomaly score of the sample $\mathbf{x}_i$ is then defined as the difference between the two means:

$$S_i^{\text{USDR}} \equiv z_i^{(\text{infer})} - z_i^{(\text{train})} \tag{7}$$

In other words, the anomaly score of a sample $\mathbf{x}_i$ quantifies the generalization ability of a model which includes $\mathbf{x}_i$ in its training set. If an input sample $\mathbf{x}_i$ is an anomaly, learning its characteristics is likely to add valuable information to the model. We thus expect a significantly lower residual when inferring on $\mathbf{x}_i$ if this sample is in the training set than if it is not included in the training set. On the other hand, for a normal sample $\mathbf{x}_i$, belonging to the dominant normal class, we expect no significant difference, whether the model is trained on this very input or not, due to its similarity to a large number of other samples in the full dataset $\mathcal{D}_0$. We quantify this idea by scoring each sample by its relative contribution to the generalization power of the model. In this way, our anomaly score takes into account not only the properties of the dominant normal class but exploits information from the abnormal samples as well.

3.2. Baselines

To evaluate the performance of the proposed USDR framework we compare the results with two generic approaches.

Blind Training. As a naive baseline we use the standard approach in the absence of any data refinement method. We train the residual-based model f "blindly" on the entire training data, despite the anomaly contamination. We then use the "blindly" trained model to infer on the same data, and obtain scores that should be used to separate normal from abnormal samples. The anomaly score for sample $\mathbf{x}_i$ is its absolute prediction residual:

$$r_i = |\mathbf{y}_i - \hat{\mathbf{y}}_i| \tag{8}$$

Since the model f should typically be trained on normal data only, its AD performance is expected to be poor when it is trained on contaminated data. The comparison with blind

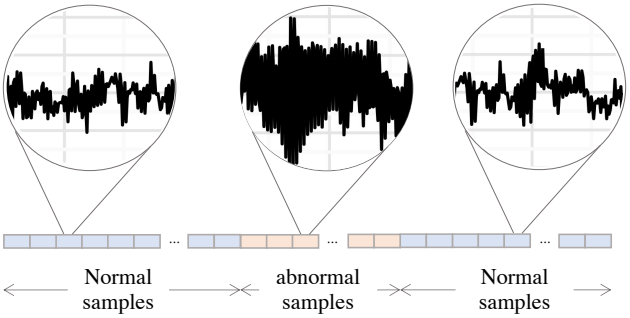

Figure 2. MIMII experiment design.

training serves to assess the improvement obtained by the refinement step we propose. In order to isolate the effect of our suggested refinement method from potential effects of the time-ordered subset generation process, we perform the blind training procedure using the same subset split as the USDR training. The anomaly score of each sample is the ensemble mean of the normalized prediction residuals of all models trained with the respective training subsets:

$$z_i = \frac{1}{M} \sum_{j=1}^{M} r_{ij} \tag{9}$$

For the sake of method comparability, the scores are then rescaled between 0 and 1.

Clean Training. In an ideal but rather unrealistic case, the training dataset would be manually cleaned prior to training, such that it contains only normal samples. In order to obtain a reference for the difficulty of the refinement task (i.e how easy the separation between normal and abnormal samples is), it is useful to train the selected prediction model f on such an optimally cleaned data (as an assumed best case scenario), and use the trained model to infer on the entire training set $\mathcal{D}_0$. To isolate the effect of the proposed method from potential effects of the subset generation, The anomaly scores are derived in the same way as for the blind model, using Eqn. 9, however with the healthy training subsets only.

The comparison of the USDR performance with the two extreme baselines is done at the level of the refinement efficacy of all three approaches, by comparing the derived scores on the original training dataset. In this way we explicitly avoid selecting a refinement threshold, a task which would strongly depend on the selected model. The latter will be addressed in a separate study, focusing on state-of-the-art AD performance tests.

4. EXPERIMENTS

In order to demonstrate the performance of our suggested method, we conduct tests on two public datasets of technical machine data. The first dataset is the sound data for Malfunctioning Industrial Machine Investigation and Inspection (MIMII) (Purohit et al., 2019), containing acoustic signals of normal and abnormal machine components. With this data we construct a use case that mimics the abrupt appearance of faults in a previously healthy machine. The second dataset, the Turbofan engines N-CMAPSS DS02 of NASA (Saxena & Goebel, 2008) contains full degradation trajectories of aircraft engines. The two examples differ not only in the data type and physical context but also in their different fault dynamics, which is abrupt in one case vs. slowly degrading in the other.

The purpose of the evaluation below is to demonstrate the generic nature of our proposed approach, which does not depend on the dataset nor on the selected model f. The only prerequisite is that f is a residual-based model. We show that the unsupervised data refinement (USDR) step significantly improves the performance of the model f compared to the alternative of training blindly on contaminated data ("blind training"). As a reference for the separability of normal and abnormal samples, we show in each case the results of training with the anomaly-free part of the data ("clean training").

It is important to note that the goal of this work is not to obtain state-of-the-art AD performance on the analyzed datasets, but rather to demonstrate the benefit achieved by applying the refinement framework, independent of the chosen prediction model and of the dataset. Therefore, we do not perform extensive studies to select the best performing model, but rather demonstrate the performance on two of the most popular AD models, PCA and fully connected Autoencoder (AE). The model parameters are minimally adjusted to achieve good performance for the ideal "Clean Training" case. Since the framework is model agnostic, we expect a similar performance enhancement due to the data refinement step, if one replaces these basic standard models with state of the art alternatives (e.g more complex, including generative AE variants), depending on the use case at hand.

4.1. Acoustic Signal Data

Description of the dataset. The MIMII dataset (Purohit et al., 2019) contains audio recordings from four types of machines: fans, pumps, valves and slide rails. Some of the recordings were taken when the machines were normally functioning and others in malfunctioning states. For each machine type, normal and abnormal data from 4 individual units is recorded, with no further labels of the fault type of operating condition. For each unit, data under 3 signal-to-noise (SNR) levels (6, 0 and -6dB), controlling the background noise (unrelated to the machine functioning state). Similar to the orig-

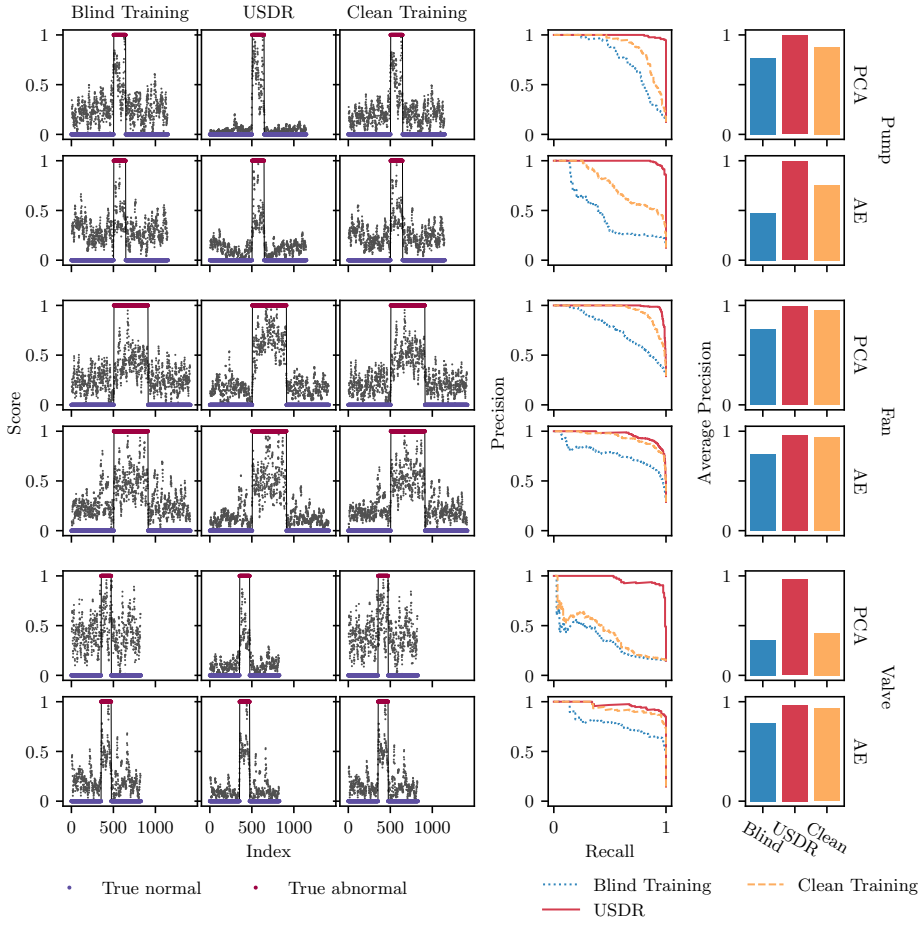

Figure 3. Examples of the framework performance for the MIMII data (test case I). The derived scores of the USDR framework are compared with the scores of blind training with the contaminated data and to clean training with normal data as a reference, calculated using PCA (upper row) and AE (lower row). The results are shown for selected cases: Pump (id00,0dB), fan (id00,6dB), and valve (id02,0dB). The two columns on the right show the precision-recall curves (PRC) and the average precision (AP) for the three methods.

inal paper (Purohit et al., 2019), we use only the first channel of microphones, and consider log-Mel spectrograms as the input feature to the AD models. The spectrograms are generated as described in the original dataset description (Purohit et al., 2019).

Experiment design. We use the MIMII dataset to mimic a realistic industrial setting in which continuously recorded condition monitoring data from a given machine is contaminated with unlabeled faults. To this end, we concatenate all of the acoustic signals from the MIMII dataset which belong to a single individual unit (under fixed SNR conditions). We would like to approximate a situation in which a machine is turning faulty (malfunctioning) in an abrupt way, and is then being repaired thus regaining its normal functioning. This results in a time series that contains a first normal (healthy) period (constructed by concatenating n_0 acoustic signals), then one (or more) faulty periods followed by normal periods due to the repair. Each normal (abnormal) period is constructed by signal concatenation of n_h (n_f) acoustic signals. An example of how the time series of concatenated signals is constructed is shown in Figure 2. We note that the faulty period may contain a mixture of failure modes, as the original dataset does not contain labels of the fault types or root causes.

Test cases. We consider two test cases, mimicking two different anomaly structures in the MIMII training data. In test case I, we construct a condition monitoring signal with a single faulty period, and in test case II there are 3 short faulty periods containing abnormal signals. The purpose of this is to

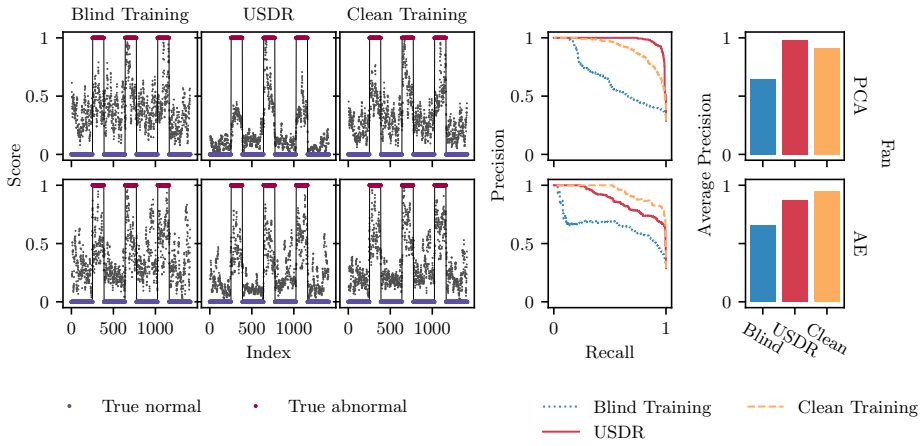

Figure 4. Examples of the framework performance for the MIMII data (test case II). The figure structure is similar to Fig. 3, demonstrated for the Fan system. Here we assumed three short faulty periods instead of a single long fault.

demonstrate the effectiveness of the framework for different contextual anomalies, approaching the limit of point anomalies. The latter will be further tested in a separate study.

Within each test case we use the USDR framework to refine the entire data available in the dataset for a specific unit. We repeat this for all units of the 4 machine types: fan, pump, slide rail, and valve.

To this end, the available data for a given unit is split using a sliding window of a fixed length of 200 samples into partially overlapping training sets, such that each sample repeats in $M_{\text{train}} = 5$ sets.

Results. Selected results of test case I are shown in Figure 3. The three columns on the left show the derived scores for the three methods: blind training, USDR (the proposed framework), and clean training. The scores are shown for three different units of three machine types (fan, pump, and valve) with two simple reconstruction models, PCA and fully connected AE, for each machine type. In all cases we trained a PCA with 5 principal components and an AE with 7 dense layers, a latent dimension of 80 (resulting in 236944 trainable parameters) and ReLU activations. In each panel, the true labels are marked in color for reference: blue for normal samples with label 0 and red for abnormal ones with label 1. The displayed scores have a different derivation for each of the three methods. For Blind Training and Clean Training, we show the anomaly scores obtained by rescaling the reconstruction errors (Eqn. 8 with $\mathbf{y}_i = \mathbf{x}_i$ and $\hat{\mathbf{y}}_i = \hat{\mathbf{x}}_i$) to range between 0 and 1. For the USDR framework we show the refinement scores S_i^{USDR} obtained using Eqn. 7. To improve the separability all scores are smoothed with a moving mean of 10 samples. We note that this step can be avoided by further optimizing the choice of the prediction model f to ideally fit the use case at hand, a step that we did not undertake here in order to remain simple and emphasize the model independence of our framework.

For each machine unit and SNR condition, we use all available normal and abnormal samples as the initial training data with the aim of data refinement, i.e. selecting only the normal samples in order to use them to retrain an AD model in the next step. Note that we do not show the results of such a second step. In this way the results we show remain generic, with no need to determine a refinement threshold. In contrast to our proposed framework, the optimal threshold would be model-dependent.

In addition to showing the derived refinement scores, Figure 3 shows the corresponding precision-recall curves (prc) and their resulting average precision (AP) on the two right-most columns. The prc curves result from evaluating the scores obtained by each of the three methods (blind, USDR, and clean) against the true normal/abnormal labels of the training samples. In this way, the prc measures the quality of the separation obtained by each method between normal and abnormal samples in the contaminated dataset. The corresponding AP values are displayed as bars in the right-most panel of each row.

From Figure 3, it is evident that the Blind Training (left most column) performs poorly for all machine types and with both PCA and AE models. In many cases, it is close to impossible to find a threshold that would lead to a proper refinement of the training data based on the blind anomaly scores.

The suggested USDR framework (second column from the left), however, shows a clear separation between the refinement scores of true normal and true abnormal samples. This allows to select a threshold that removes most of the abnormal

Table 1. Average precision (AP) scores for AD on contaminated MIMII data.

			PCA			AE			OC-SVM	IF
	Contamination		Blind	USDR	Clean	Blind	USDR	Clean		
Fan	29%	6 dB	0.94	1.00	0.99	0.90	0.98	0.97	0.86	0.97
		0 dB	0.75	0.93	0.85	0.73	0.83	0.83	0.60	0.77
		-6 dB	0.55	0.67	0.60	0.50	0.63	0.64	0.37	0.36
Pump	12%	6 dB	0.95	1.00	0.98	0.96	1.00	0.99	0.61	0.90
		0 dB	0.71	0.99	0.84	0.65	0.91	0.83	0.5	0.67
		-6 dB	0.39	0.77	0.45	0.41	0.58	0.52	0.22	0.41
Slider	25%	6 dB	0.77	0.98	0.95	0.82	0.91	1.00	0.46	0.80
		0 dB	0.60	0.94	0.77	0.88	0.96	0.99	0.38	0.75
		-6 dB	0.35	0.60	0.50	0.72	0.80	0.87	0.23	0.32
Valve	11%	6 dB	0.41	0.66	0.49	0.79	0.92	0.90	0.31	0.46
		0 dB	0.20	0.47	0.24	0.56	0.70	0.71	0.24	0.31
		-6 dB	0.12	0.14	0.12	0.27	0.48	0.43	0.13	0.09

samples without removing many normal samples from the training data, thereby refining the data for subsequent training of a residual-based AD model. The separability obtained by the USDR is comparable to the Clean Training reference (third column from the left), and is often higher. The comparison of the three methods can be quantified in terms of the prc results on the two columns on the right. The USDR method (red) always significantly outperforms the blind training (blue) in the achieved AP, and performs at least as well as the ideal reference trained with anomaly-free data (yellow). In several cases it is seen to outperform the clean reference, which may be interpreted as the "best-case-scenario". This result can be intuitively understood when considering that the USDR method explicitly contrasts the information from abnormal samples with the information from normal ones, by quantifying the contribution of each sample to the generalization ability of the trained model at inference time.

We note that depending on the machine type, either PCA or AE show better refinement performance. A similar model dependence is also observed for the Clean Training anomaly scores displayed in the third column for reference. For example: even in the ideal case of anomaly-free training data, PCA performs quite poorly on the valve data, whereas AE displays a rather high separability between normal and abnormal samples. It is interesting to note, that in the PCA case the benefit of the USDR is particularly high. In this case, the residuals themselves contain little information about the anomaly. However, the drop in the residual of a specific sample once it is included in the training set is the discriminating factor of anomalies from normal samples.

The results of this test case are summarized in Table 1. For each machine type and SNR, the mean score over ten repetitions and the four units is shown. This is repeated for the three methods (blind training, USDR and clean training) using both PCA and AE. As a reference for the task complexity, we display the mean score achieved by two simple unsupervised baselines: one-class support vector machine (OC-SVM) and isolation forest (IF). The performance of these two naive approaches deteriorates as the task of separation between the normal and abnormal samples gets more difficult. We note that the OC-SVM baseline was calculated assuming that the anomaly fraction is known (as opposed to the proposed USDR which does not assume this). The table demonstrates a broad range of difficulties of the AD tasks, with various contamination ratios and different fault features for each machine type and SNR level. Accordingly, the relative performance of different AD models can strongly vary. However, the common results to almost all cases is a significant improvement achieved by the USDR method compared to the blind training when both are using the same underlying model (either PCA or AE). The refinement task becomes harder when the SNR ratio decreases. Particularly noisy cases (e.g. Valve with -6dB) are hard to refine, also for the clean training reference model. Another common outcome is the fact that the USDR is on par with the ideal Clean Training case (which would not be available in many real-world applications), and sometimes outperforms it, given a selected model f (here PCA or AE).

We note again, that the purpose of the present study is not to obtain state-of-the-art AD performance, but rather to demonstrate the striking improvement achieved by the suggested refinement framework, compared to the alternative of training with contaminated data. This improvement is shown here across machine types, contamination ratios, and prediction

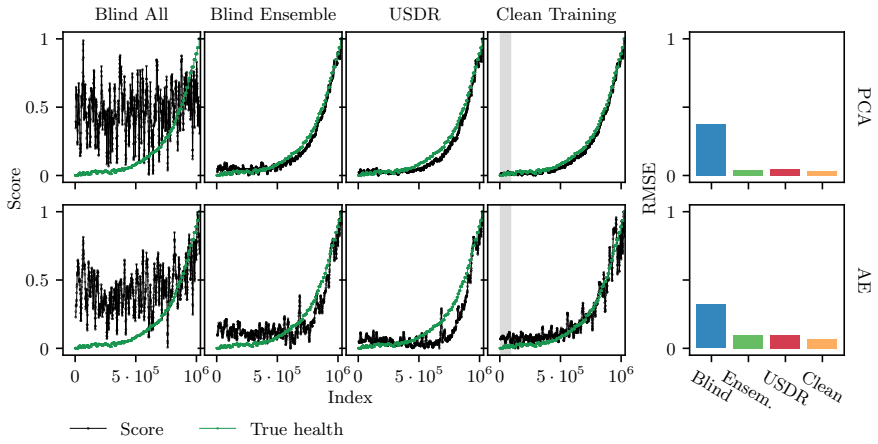

Figure 5. Examples for the framework performance with the turbofan engine CMAPSS data. The derived scores of the USDR framework are compared with the scores of blind training with the all contaminated data, blind ensemble, and clean training with normal data (first 10 engine cycles) as a reference, calculated using PCA (upper row) and AE (lower row). The results are shown for engine 5. The column on the right shows the RMSE for the four methods.

models. The absolute refinement performance can be further improved by optimizing the prediction model at hand, which is beyond the scope of the present work.

Exemplary results from Test case II are displayed in Figure 4. The structure of the figure is identical to the one of Figure 3 and the results are shown for the Fan, containing 29% contamination in the training data. The two cases differ in the way the training samples were concatenated to construct the condition monitoring time series. In test case II we mimic a situation of three short faulty periods instead of a single longer fault. The purpose of this additional test is to demonstrate the performance improvement achieved by the USDR framework, for several different abrupt fault scenarios. The next step in this direction, of addressing the case of completely isolated outliers, is part of our future research.

Figure 4 demonstrates again that the proposed USDR method significantly outperforms the blind training alternative, and often reaches the performance of the unrealistic Clean Training scenario.

4.2. Aircraft Engine Data

Description of the dataset. As a second use case, we evaluate the performance of the proposed framework on the well-known turbofan engine dataset. This dataset contains synthetic run-to-failure degradation trajectories of nine aircraft engines, generated with the Commercial Modular Aero Propulsion System Simulation (CMAPSS) model of NASA (Frederick, DeCastro, & Litt, 2007), and was generated taking real flight conditions from commercial jets as input. The dataset includes 19 variables of the flight conditions as well as the temperatures and pressure levels at various parts of the engine, for multiple flight cycles, from the beginning of life until full degradation of the engine. Different engines (units) display different failure modes, some affecting a single component and others affecting multiple components within the engine (Chao, Kulkarni, Goebel, & Fink, 2022). A standard training procedure with the CMAPSS data uses a residual-based model (often a reconstruction model with an AE architecture) with all 19 time series variables as input and output. In the standard approach, the first flight cycles of a given engine are assumed to represent healthy behavior and are thus used to train the reconstruction model. At inference time, the model is expected to yield large reconstruction errors as the engine condition deteriorates, that is with a growing cycle number. Various extensions of this basic architecture were suggested along the years in order to predict the remaining useful life (RUL) of the various unit engines (see for example (Li, Ding, & Sun, 2018; Arias Chao, Kulkarni, Goebel, & Fink, 2021). Here, however, we do not aim to predict the RUL but use this dataset for the purpose of unsupervised detection of abnormal behavior, observed in this case as a slow degradation of the condition of the engine. In this setting, any deviation from the normal behavior is regarded as "anomaly" with an assigned anomaly score, which ideally reflects the degree of degradation and can thus be converted into a health index. At a later stage, which is beyond the scope of the present paper, the health index may be used to predict the RUL.

Results. Figure 5 displays the refinement scores derived using the USDR framework for the entire degradation trajectory of a single engine, in this case unit 5 of the CMAPSS DS02 dataset. As in the MIMII use-case, the available data for this unit is split using a sliding window of length 20% of the avail-

Table 2. RMSE of Anomaly Scores for Turbofan Data

Unit	PCA				AE			
	Blind all	Ensemble	USDR	Clean	Blind all	Ensemble	USDR	Clean
2	0.34	0.06	0.07	0.03	0.29	0.15	0.13	0.07
5	0.38	0.04	0.04	0.02	0.32	0.10	0.10	0.05
10	0.33	0.04	0.05	0.03	0.31	0.12	0.11	0.08

able data into $M = 20$ partially overlapping training subsets, such that each sample repeats in $M_{\text{train}} = 4$ subsets. Here as well, we contrast the refinement scores of the USDR with the anomaly scores (rescaled reconstruction errors) of the Blind Training approach and the Clean Training for reference. In this case we use "Blind All" to refer to a single training with the entire data set, and "Blind Ensemble" to refer to training an ensemble of models, for which the ensemble mean of the normalized residuals is used as a score. The Clean Training scores are obtained by training the reconstruction models with data from the first 10 engine cycles (which are assumed to be degradation-free and are marked with a grey background) and inferring on the full degradation data. In each case we trained two simple residual-based models, PCA and a fully connected AE. For the PCA we select 15 principal components, and the AE is trained with 5 dense layers, a latent dimension of 20 (resulting in 6908 trainable parameters), and ReLU activations.

In all cases we rescale the scores between 0 and 1 and compare them with the true health index provided for this dataset (green). The left column of Figure 5 shows clearly that blindly training a reconstruction model using the full degradation trajectory with no refinement fails to reveal the true health condition of the unit. However, both the Blind Ensemble (second column from the left) and the USDR score (third column from the left) follow the true health index rather accurately, in particular using a simple PCA model, allowing to clearly separate healthy from degraded data. Indeed, as shown in the rightmost column, the RMSE of these two scores with respect to the true health index is almost as small as the one of the Clean Training reference (fourth column from the left). It is worth noting that for this dataset, the USDR and the simpler ensemble refinement are similarly powerful in discovering the health condition of the engine in a fully unsupervised manner. They reach remarkably low RMSEs considering the high data contamination: only around 20% of the entire training data can be considered completely "normal". The success of both ensemble-based approaches to discover the degradation pattern potentially hinges on the similarity between healthy and degraded conditions (as we know that only very subtle degradation effects were injected in the simulation), except for towards the very end of the degradation trajectory.

Table 2 shows the AD performance for three different engines. In order to compare with a clear ground truth health indicator we selected the three engines with a simple degradation mode, of an abnormal high pressure turbine (HPT) efficiency degradation, for which the true health index can be easily extracted. The Table displays the RMSE of the different scores (Blind All, Blind Ensemble, USDR and clean training) with respect to the true health index, each with both PCA and AE as a reconstruction model (where the mean over 10 repetitions is shown). A similar conclusion follows for all three units: given a reconstruction model (PCA or AE), both refinement frameworks allow to identify the normal and abnormal segments of the data in a fully unsupervised way, achieving an RMSE which is only sightly higher than the ideal reference (anomaly-free training data). The improvement over training blindly with the entire data is striking, reducing the RMSE from around 0.3 or 0.4 down to around 0.05.

5. CONCLUSIONS

In this paper we suggest a novel data centric approach to deal with the challenge of AD without any labels, and with potentially contaminated training data. The problem is highly relevant in real-world scenarios, where labeling is expensive and sometimes impractical, and where mislabeling is a common issue. The proposed USDR framework avoids assuming that the training data is anomaly-free but rather allows for an unknown fraction of anomalies, possibly of various types and severities. It suggests a fully-unsupervised refinement of the training data, based on training any residual-based (reconstruction or regression) model on partially overlapping subsets of the contaminated training set. A refinement score is derived for each training sample based on its contribution to the generalization ability of the model, which is shown to increase significantly for abnormal samples. The advantage of the proposed approach lies in its simplicity and generic data-centric nature; it is model agnostic and is conceptually applicable to any data type. It can be applied either to the raw samples or to their learnable representations. We demonstrate the refinement efficacy of USDR for AD of contextual anomalies in multivariate time-series data from industrial machines (contamination fraction up to 29%) and aircraft engines (with only 20% normal data). We show that it performs similarly and sometimes better than a model trained on anomaly-free data. In this paper we focus on demonstrating the generic character of the framework, leaving the comparison with state-of-the-art AD models to future work. Ex-

tending the applicability of the proposed method to other data modalities (e.g. image data) and testing its hyperparameter sensitivity are additional topics for future research.

REFERENCES

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 3395–3404).

Beggel, L., Pfeiffer, M., & Bischl, B. (2020). Robust anomaly detection in images using adversarial autoencoders. In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2019, würzburg, germany, september 16–20, 2019, proceedings, part i* (pp. 206–222).

Berg, A., Ahlberg, J., & Felsberg, M. (2019). Unsupervised learning of anomaly detection from contaminated image data using simultaneous encoder training. *arXiv preprint arXiv:1905.11034*.

Bergman, L., & Hoshen, Y. (2020). Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Frederick, D. K., DeCastro, J. A., & Litt, J. S. (2007). *User's guide for the commercial modular aero-propulsion system simulation (c-mapss)* (Tech. Rep.).

Golan, I., & El-Yaniv, R. (2018). Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, *31*.

Hendrycks, D., Mazeika, M., & Dietterich, T. (2018). Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Latecki, L. J., Lazarevic, A., & Pokrajac, D. (2007). Outlier detection with kernel density functions. In *Mldm* (Vol. 7, pp. 61–75).

Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, *172*, 1–11.

Michau, G., Frusque, G., & Fink, O. (2022). Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences*, *119*(8), e2106598119.

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). Deepant: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access*, *7*, 1991–2005.

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., & Kawaguchi, Y. (2019). Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv:1909.09347*.

Qiu, C., Li, A., Kloft, M., Rudolph, M., & Mandt, S. (2022). Latent outlier exposure for anomaly detection with contaminated data. In *International conference on machine learning* (pp. 18153–18167).

Saxena, A., & Goebel, K. (2008). *Turbofan engine degradation simulation data set.* NASA Prognostics Data Repository. Moffett Field, CA.

Schneider, T., Qiu, C., Kloft, M., Latif, D. A., Staab, S., Mandt, S., & Rudolph, M. (2022). Detecting anomalies within time series using local neural transformations. *arXiv preprint arXiv:2202.03944*.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. *Advances in neural information processing systems*, *12*.

Shenkar, T., & Wolf, L. (2022). Anomaly detection for tabular data with internal contrastive learning. In *International conference on learning representations.*

Sohn, K., Li, C.-L., Yoon, J., Jin, M., & Pfister, T. (2020). Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*.

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*, 45–66.

Wang, G., Zhan, Y., Wang, X., Song, M., & Nahrstedt, K. (2022). Hierarchical semi-supervised contrastive learning for contamination-resistant anomaly detection. In *European conference on computer vision* (pp. 110–128).

Yoon, J., Sohn, K., Li, C.-L., Arik, S. O., Lee, C.-Y., & Pfister, T. (2021). Self-supervise, refine, repeat: Improving unsupervised anomaly detection. *arXiv preprint arXiv:2106.06115*.

Zgraggen, J., Guo, Y., Notaristefano, A., & Goren Huber, L. (2023). Fully unsupervised fault detection in solar power plants using physics-informed deep learning. In *33rd european safety and reliability conference (esrel), southampton, united kingdom, 3-7 september 2023* (pp. 1737–1745).

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., ... Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 1409–1416).

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd acm sigkdd international conference on knowl-*

edge discovery and data mining (pp. 665–674).

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations.*